

THE PROCESS OF DEVELOPING A LEXICAL DATABASE USING TECHNOLOGY

Sharipova Aziza Abdumannapovna

Tashkent University of Information Technologies named after Muhammad al-Khwarizmi, Foreign Languages Department, PhD., ass.prof.

ARTICLE INFO.

Key words:

corpus, FLT, SLT, meta-analysis, dictionary, technology, vocabulary corpus.

Annotation

With the rapid development of information technology, researchers worldwide strive to integrate IT closely into their professional field. Being a part of this modernity, linguists try to focus on the advantages of information technology in terms of promoting the teaching/learning of foreign languages. This article explores the potential benefits and approaches to using corpus databases in vocabulary learning in foreign language lessons. It is believed that overall language proficiency largely depends on competence in knowing and using words in their true context (Carter and McCarthy, 1988). However, almost all learners studying English as a second or foreign language (ESL/EFL) encounter some difficulties with vocabulary, which becomes one of the main challenges in language learning (Cobb, 2003). On the other hand, it cannot be ignored that vocabulary is an integral part of language and plays a crucial role for learners studying English as English (Zhang & Liu, 2014).

<http://www.gospodarkainnowacje.pl/> © 2023 LWAB.

Historical background and definition of corpus in linguistics

The most suitable definition of the word "corpus" was given by Rappaport in 2010, saying that a corpus is a large principled collection of natural language stored in electronic form. It can be stated that a corpus is a complex collection of authentic oral and written words, systematically compiled according to certain principles and presented in an electronic database. Any genre of texts can be gathered in a corpus, including newspapers, Facebook posts, recipes, novels, speeches, scripts, conversations with friends, letters, books, magazines, lectures, essays, notes, etc. On the other hand, the first records of the use of corpora in linguistics can be traced back to 1964, when the initially mechanized corpus "The first Brown Corpus" was created. As for Meyer's assumptions, corpus etymologists are largely intrigued by graphical or helpful clarifications of language and study etymological wonders through observing substantial mechanized language databases called corpora. In 1998, Biber, Conrad, and Reppen defined a corpus as a "large and principled collection of natural texts" constituted as a representation of language as a whole, a language, or another subset of language. Corpora may contain language in terms of compositions, deciphered discourse, or both (Conrad, S. 2005). Electronic versions of all the above-mentioned were subsequently removed by being dissected with the help of PC software projects called concordancers or, not exclusively, concordancing programming. Since corpus linguistics was approved

as a method, gathered content testified that the ultimate goal of linguistics wonders analysis was not an entirely new idea. Once pointing to Meyer's late article, early references to words relied on a sizable collection of distributed works and a vast number of references to actually existing language.

Advantages of a corpus

According to research conducted by Kousha and Jafarpour (2006), it was noted that the use of corpus-based materials tends to be highly effective in terms of teaching prepositional collocations. As for the work of Huang (2014), it was found that the use of corpus-oriented DDL activities to improve the lexical-grammatical use of abstract nouns in L2 writing was successful. The subsequent results showed that learners were able to demonstrate greater variety of collocations and phrases in their writing practices, avoiding linguistic errors when using target abstract nouns. As for another study conducted by McArthur and Littlemore (2008), which investigated the usefulness of corpus data for learners in understanding the definition of nominal verbs in English and Spanish.

A corpus, which is a collection of written or spoken texts, has several advantages in research and language analysis. Some of the major advantages of using a corpus are:

1. **Authenticity:** A corpus consists of real language use, providing authentic and natural examples of how a language is used in different contexts. This authenticity helps researchers better understand the nuances, variations, and complexities of the language.
2. **Representative sampling:** Corpora are carefully sampled to represent different genres, registers, time periods, and other relevant factors. This representative sampling ensures that the analysis and findings reflect a broad range of language usage.
3. **Quantitative analysis:** Corpora allow for quantitative analysis by providing large amounts of data. Researchers can analyze frequency patterns, collocations, and other statistical measures to understand language patterns and tendencies. This quantitative analysis provides objective evidence for language phenomena.
4. **Detailed examination:** Corpora enable researchers to conduct detailed examinations of language features, such as grammar, vocabulary, and discourse. By analyzing real language data, researchers can identify and study specific linguistic patterns, structures, and meanings in depth.
5. **Comparative analysis:** Corpora facilitate comparative analyses by allowing researchers to compare different varieties of a language, different time periods, or different genres, among other factors. This comparative analysis helps identify similarities and differences, contributing to a deeper understanding of language variation and change.
6. **Corpus-driven research:** Corpora can drive research by providing evidence-based insights and generating new research questions. Researchers can explore patterns in the data, derive hypotheses, and propose new theories or models based on the findings from corpus analysis.
7. **Language teaching and learning:** Corpora have practical applications in language teaching and learning. They can be used to develop teaching materials, create language exercises, and assist in vocabulary acquisition and language proficiency development. Learners can also benefit from studying authentic language use in corpora to improve their comprehension and production skills.

Definition of vocabulary

A convenient description of vocabulary can be found in the guide "Corpus Linguistics for Vocabulary" written by Pavel Shudarsky, which describes it as a fundamental element of language use, and research investigating its relationship with other components of linguistic competence constitutes an important area of applied linguistics. In this section, the author describes the role of vocabulary as a factor influencing overall language proficiency and as a key element in approaching native-like fluency. Collocation of words is another aspect of vocabulary where corpora can be used as a source of important insights. Various lexical relations of words demonstrate that they do not exist in isolation and can be studied in different lexical contexts. Synonymous words can be considered as examples of the aforementioned lexical relations along with metaphorical expressions, which are also studied using corpus analysis. Furthermore, idioms or idiomatic expressions, described as a special kind of non-literal

language, also attract linguists' attention due to the applicable ideas that can be utilized when analyzing corpus-based lexical variations that arise from the use of certain phrases in different contexts and social conditions.

Approaches to corpus-based language learning for vocabulary

Database approach

A single definition of corpus and vocabulary is not enough for successful integration of knowledge into the educational system. Therefore, this part of the article discusses possible methods of corpus integration in vocabulary learning. The primary tool of the corpus that can be utilized in vocabulary TFL is corpus-informed analysis. This tool helps learners obtain substantial data about language usage in real conversations. As for Partridge, a corpus is a collection of authentic texts (written or transcribed oral), stored in electronic form [9, p. 103]. The magnitude of vocabulary within can vary from a few semantic structures to millions of words. As mentioned earlier, linguistic information is usually represented in the form of concordances (Tribble & Jones, 1997). Generating concordances is possible using the aforementioned software, Concordance. In terms of previous research, Tim Jones was noted as one of the first teachers to incorporate concordancing and became the author of data-driven learning (Jones, 1991). DDL (data-driven learning) can be seen as a language learning method based on the hypothesis that authentic language input along with concordancing allows learners to explore language as it is used in real-life situations. Another important fact is that in DDL, the language acquisition process is focused on the discovery of linguistic rules and usage patterns by learners.

Collaboration of foreign linguists (Deriabina I.V., Klochikhin V.V., Kokoreva A.A., Sysoev P.V., Cherniakova T.A., Gabrielaitis K., Lakman K., McCarthy M., Sinclair J., etc.) has put forth the idea of developing learners' lexical skills using an electronic linguistic corpus. In his research, P.V. Sysoev defines a linguistic corpus as a collection of texts that are united into a single system based on certain characteristics (language, genre, time of creation, author, etc.) and equipped with a search engine [6, p.100]. The didactic potential of a corpus of texts in teaching ESL vocabulary lies in the opportunity to use various authentic examples of language usage that vividly demonstrate the functioning of linguistic and speech units, collocations, idioms, etc. [7, p.70]. According to the research, the main characteristics of a linguistic corpus that enable its use in the development of lexical skills include diversity of functional text types, context-dependent search results, and relevance of textual data. Working with a linguistic corpus in ESL classes contributes to the formation of language intuition, acquisition of collocations, understanding of the peculiarities of lexical unit functioning in speech, and so on.

Conclusion

Learning a foreign language has become an integral part of the basic learning process for individuals. The aim of this article was to find applicable methods of corpus integration in TFL vocabulary classes to enhance the potential for language learning progress. The article explores historical prerequisites for the connection between corpus and linguistics, identifies collocations as a corpus, concordances, corpus linguistics, and vocabulary, and examines two approaches to corpus integration in the process of vocabulary learning. It is appropriate to emphasize that the field of corpus-lexicography connection requires further research and analysis.

References:

1. Edel, A. (2010). Using Corpus Linguistics in the Study of Political Discourse, in A. O'Keeffe and M. McCarthy (eds.) *The Routledge Handbook of Corpus Linguistics*. London and New York: Routledge.
2. Pavel, S. (2017). *Corpus Linguistics for Lexicography: A Guide for Research*. London: Routledge.
3. Aravind, B., & Rajasekaran, V. (2020). Qualitative Research Using a Novel Technique to Enhance Vocabulary for Students Learning English as a Foreign Language. *International Journal of Emerging Technologies*, 11(2).

4. Krylova, E., Utkina, N., & Wang, S. (2020). On Modern Methods of Vocabulary Teaching. Innovative Methods of Teaching and Upbringing. Collection of articles from the 2nd International Scientific and Practical Conference. Penza: Nauka i Prosveshchenie.
5. Meyer, C. F. (2002). English Corpus Linguistics: An Introduction. Cambridge: Cambridge University Press. <http://dx.doi.org/10.1017/CBO9780511606311>
6. Biber, D., Conrad, S., & Reppen, R. (1998). Corpus Linguistics: Investigating Language Structure and Use. Cambridge: Cambridge University Press. <http://dx.doi.org/10.1017/CBO9780511804489>
7. Conrad, S. (2005). Corpus Linguistics and L2 Teaching. In E. Hinkel (Ed.), Handbook of Research in Second Language Teaching and Learning (pp. 393-409). Mahwah, New Jersey: Lawrence Erlbaum Associates.
8. Ibragimova, N.A. "Language In The System of Artificial Intelligence" International Conference on Information Science and Communications 2019.
9. Partridge "Rawls and the Duty to Posterity. A Doctoral Dissertation. The University of Utah", 1976.